HEALTH EDUCATION RESEARCH
Theory & Practice

Vol.21 (Supplement 1) 2006
Pages i47–i57
Advance Access publication 31 October 2006

Item response modeling: an evaluation of the children's fruit and vegetable self-efficacy questionnaire

Kathy Watson*, Tom Baranowski and Debbe Thompson

Abstract

Perceived self-efficacy (SE) for eating fruit and vegetables (FV) is a key variable mediating FV change in interventions. This study applies item response modeling (IRM) to a fruit, juice and vegetable self-efficacy questionnaire (FVSEQ) previously validated with classical test theory (CTT) procedures. The 24-item (five-point Likert scale) FVSEQ was administered to 1578 fourth graders from 26 Houston schools. The IRM partial credit model indicated the five-point response options were not fully utilized. The questionnaire exhibited acceptable (>0.70) reliability except at the extremes of the SE scale. Differential item functioning (DIF) analyses revealed no response bias due to gender. However, DIF was detected by ethnic groups in 10 items. IRM of this scale expanded what was known from CTT methods in three ways: (i) areas of the scale were identified that were not as reliable, (ii) limitations were found in the response format and (c) areas of the SE scale levels were not measured. The FVSEQ can be improved by including items at the extreme levels of difficulty. DIF analyses identified areas where IRM can be useful to improve the functioning of measures.

USDA/ARS, Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, 1100 Bates Street, Houston, TX 77030, USA
***Correspondence to: K. Watson.**
E-mail: kwatson@bcm.tmc.edu

Background

Low intake of fruit and vegetables (FV) is a health risk behavior underlying the leading causes of mortality and morbidity in the United States [1–3]. FV intake practices typically develop during childhood, thereby increasing the need to encourage FV intake early in life. Perceived self-efficacy (SE) [4] for eating FV is a key variable mediating change from intervention programs and thereby is critical to effective strategies for interventions [5, 6]. Several studies [7–14] have used original or modified versions of the FV SE scales developed by Domel *et al.* [15] and for the CATCH study [16], while other studies [17–19] have used a modified version of the FV SE instrument for adults [20] or developed their own [5, 21, 22]. The scales vary in the number of subscales, the number of items and the response options with one study [23] using only a single item to measure FV SE. Nearly all of the studies have demonstrated adequate internal consistency reliability ($\alpha > 0.70$). Test–retest reliability examined in a few of the studies has been at least modest ($\alpha > 0.60$). Two studies [8, 12] showed significant increases in SE over time. Two studies [7, 13] reported conflicting results when comparing intervention and control groups. Several studies reported significant relationships between SE and FV intake/change in intake [10, 11, 14, 19, 24], fruit intake only [25], FV snack choice [17] and sugar [21]. Three studies [9, 21, 22] reported no significant association between SE and FV intake/intake-related behaviors. In studies that compared modes of assessment (computer versus paper), one study

doi:10.1093/her/cyl136

[23] reported a significant difference while another study did not [18].

Using classical test theory (CTT) techniques [26], the FV self-efficacy questionnaire (FVSEQ) used in this study exhibited acceptable internal consistency reliability (Cronbach's α between 0.72 and 0.90) and test–retest reliability (0.70) [7, 11, 15]. Construct validity was determined through principal components analyses which yielded acceptable loadings (> 0.40) on two subscales [7, 11]. Although a couple of studies have reported gender [27] and ethnic [28] differences in psychosocial correlates of diet in children/adolescents, no one has analyzed the item functioning, response format or ethnic or gender differences in item functioning of such a scale among children.

Item response modeling (IRM) is a step beyond CTT and links a person's latent ability to the probability of selecting a specific response [29–34]. IRM can determine the best response format of an item from data, the amount of information provided by an item, the fit of a latent model to a set of items with deviations representing measurement error, the reliability of a test across the continuum of the ability [35] and group-related differences in item functioning (DIF) [36]. Anticipating the increased need to more precisely measure fruit and vegetable self-efficacy (FVSE) in children, the aim of this study was to evaluate its psychometric properties and to investigate differences in the psychometric properties across gender and ethnic groups using IRM [29, 30, 36, 37].

Methods

Sample

The sample included 1578 fourth-grade students from 26 elementary schools in the Houston Independent School District recruited to participate in the baseline assessment of the 'Squire's Quest' program evaluation [38]. The schools were randomly assigned to the treatment and control groups and all fourth graders were invited to participate. Informed assent from students and consent from parents to participate were received from 73.2% of the students in the treatment group and 67.6% of the students in the control group. The Squire's Quest program implemented in the treatment schools was a psychoeducational interactive multimedia game delivered in 10 sessions over a 5-week period. Each session lasted ~25 min. The program used social cognitive theory to attempt to increase (i) fruit, juice and vegetable (FJV) preferences, (ii) asking behaviors for FJV and (iii) skills in FJV preparation. The program also attempted to associate fun with consumption of FJV. The Institutional Review Boards of both the University of Texas MD Anderson Cancer Center and the Baylor College of Medicine approved the study protocol.

Instruments

The FVSEQ

The FV SE items [11, 15] were assessed with 24 items and included two subscales: shopping/asking SE and selection SE. The original FV SE subscales [15] employed 34 items with a three-point item format: 'not at all sure', 'a little sure' and 'very sure'. The original scale had high internal consistency (Cronbach's $\alpha = 0.88$ and 0.92) and adequate 2-week test–retest reliability ($r = 0.70$). In the original study, content validity of the instrument in the original study was assessed through a series of steps: pilot testing, revisions and principal components with stable loadings across two split-half samples. Construct validity was assessed through correlations among the FVSE subscales, FV consumption, preferences and outcomes expectancies. One subscale ('breakfast and lunch FV and paying for FV') was significantly correlated ($r = 0.18$) with FV consumption and three of four subscales were significantly correlated with preferences ($r = 0.18$ –0.49) and health/physical activity outcomes expectancies ($r = 0.13$ –0.25). In the second study [11], the scale was modified by reducing the number of items. Principal components analysis was used to assess the content validity of the reduced scale and two factors emerged. The modified scale demonstrated adequate internal consistency (Cronbach's $\alpha \geqslant 0.78$). The two subscales were significantly correlated with FV intake ($r = 0.12$) and

psychosocial characteristics (r = 0.15 –0.43). The questions in the current study were graded on a five-point Likert type scale from 'disagree a lot' to 'agree a lot' and included the same stem 'I think I can ...'. Item 1, for example, was 'I think I can write my favorite fruit or vegetable on the family's shopping list'. The internal consistency of the scale, based on Cronbach's alpha, in this sample was 0.91.

Analyses

Classical test theory

Item and test characteristics were first evaluated using CTT analysis. Item parameters included the item mean (item difficulty) and discrimination [the corrected item-total correlation (CITC)]. The internal consistency reliability of the test was calculated using Cronbach's alpha. According to Nunnally and Bernstein [39], poorly discriminating items were identified with items where the CITC was <0.30 and adequate reliability was demonstrated with a reliability index of at least 0.70. The CTT item analyses were performed using SPSS v. 11.0 [40].

Item response modeling

IRM relates the probability of selecting a given response option to a particular item as a function of the individual's level of SE [29, 30, 41]. The item parameters include measures of item difficulty (how confident the student is in their ability to eat FV) [29]. The IRM item difficulty in the dichotomous model is analogous to the item mean derived from CTT procedures. IRM individual ability estimates represent the individual's level of SE over all items and are analogous to the individual's total score in CTT.

Unidimensionality is a primary assumption for most IRM models; other issues to consider are local independence of items and adequate sample size. When unidimensionality is met, local independence may be inferred [29]. The issue of sample size for IRM (in general) and differential item functioning (DIF) analyses (specifically) is complex and depends on a number of factors (e.g. number of items, number of responses per item, the IRM model, the goal of the study, etc.). Although some researchers have suggested samples sizes of at least 500 [42] for IRM, other research has shown that smaller sample sizes may be adequate [43–45].

The assumption of unidimensionality was first tested using Principal Axis Factoring performed in SPSS [40] and assessment of percentage of variance explained. For IRM, the presence of a major factor does not necessarily preclude the presence of a minor factor. After unidimensionality was satisfied, an IRM model from the Rasch family of models was selected which best explained the data. The Rasch models assume the item slope to be fixed at 1.0 [30, 31, 41].

Two models for ordinal data were considered: the rating scale model (RSM) [46] and the partial credit model (PCM) [47]. The item difficulty issue is more complex because of the ordinal nature of the data. In addition to the overall difficulty of the item, there is difficulty associated with each subsequent response option. For example, there is difficulty associated with making the subsequent 'steps' among the response categories such as going from 'disagree a lot' to 'disagree a little' or from 'not sure' to 'agree a little'. For the RSM, the level of difficulty associated with each of the steps is the same (the category steps) across all items. This 'step' is represented as a main effect in the model. However, the PCM allows the steps to vary among the items and is represented as the item by steps interaction (item steps). The likelihood ratio test (LRT) and infit mean-square statistics (MSQ) were used to determine the best fitting model. The possible range of the MSQ fit values is between zero and infinity with values near one indicating close agreement between observed and expected values. Values <1.0 indicate less variation in the responses and values >1.0 indicate more variation in the responses than expected. Item misfit was established by infit MSQ values outside the range of 0.75 and 1.33, with significant t values [30]. The Wright map of person–item estimates displays the item difficulty and the extent to which the item difficulty reflected the full range of SE ability. Reliability was assessed, conditioned on SE ability. Visual examination of the item response functions (IRFs) assessed the functioning of the five-point Likert scale response format.

Results from the Rasch model yielded estimates for the difficulty of each item and each individual's level of SE. For the dichotomous items (0,1), the item difficulty represents the point along the SE continuum where the probability of selecting '1' is 0.5 [30, 31, 41]. For items with more than two response options, a representation of the level of difficulty can be assessed with Thurstone thresholds [30] for each of the item/category steps. The number of thresholds per item is one less than the number of options. For example, items with three response options ('none', 'a little', 'a lot') would have two Thurstone threshold estimates for the two possible steps: (i) the step from 'none' to 'a little' and (ii) the step from 'a little' to 'a lot'. Threshold 1 represents the point along the SE continuum at which the cumulative effect of the 'a little' and 'a lot' options are more likely than 'none'. Threshold 2 represents the point at which the option 'a lot' is more likely than 'none' or ' a little'. For a more complete discussion of the Thurstone thresholds, refer to Wilson, Allen and Li's [30] paper in this issue.

The estimates of difficulty in relation to person estimates are shown in the Wright item–person map. The Wright map is a figure that uses the same scale (generally in logits) to show the direct link between the distribution of participants' SE estimates (the ability) and the distribution of the items' difficulty estimates. Person estimates were obtained from the plausible values (PVs) computed during the estimation process. PVs were used, as opposed to expected a posterior estimates or maximum likelihood estimates because the PVs provided unbiased estimates of the latent abilities [48, 49]. A more in-depth discussion of the Rasch family of models may be found elsewhere in this volume [30, 31].

DIF analyses were used to identify potentially gender or ethnic group biased items [36, 43, 44, 50–52]. DIF is not represented as overall group differences but, rather, DIF concerns the expectation that participants who are in different groups but have equal levels of SE would have the same probability of selecting a particular response. Differential 'impact' (overall group differences) would be indicated by different response frequencies between groups, whereas DIF focuses on whether the participants with the same level of SE respond similarly to the item. In order to distinguish between impact and DIF, the model needs to control for differences between groups [37]. Although there are several methods for assessing DIF, in this study DIF was assessed by adding the group main effect and the group by item interaction term to the model. A significant chi-square for the group by item difficulty interaction term signified the existence of DIF; a significant group main effect signified the existence of impact. Items exhibiting significant DIF were then determined by examining the ratio of the item by group parameter estimate and its corresponding standard error; ratios >+1.96 were significant. The magnitude of DIF was determined by examining the differences in the group by item interaction parameter estimates. Because the parameters were constrained to be zero, when only two groups were considered, such as gender, this value was twice the estimate of the first group. For example, suppose hypothetically the group by item estimate for Item 1 for males was –0.44. Because the parameters were constrained to be zero, the estimate for females would then be +0.44. The difference in item difficulty (DIF) between males and females would be –0.88 (–0.44 minus 0.44). For more than two groups, DIF would be the difference in estimates between the corresponding groups. The effect of DIF was considered small (difference < 0.426), moderate (0.426 < difference < 0.638) or large (difference > 0.638) [37, 53].

If significant and meaningful DIF is found, it may indicate that the interpretation of the scale may differ by group and that the scale may be gender or cross-culturally biased. Although DIF analyses identify items that statistically function differently, heuristic examination of the items is needed to ascertain item bias [54]. Items exhibiting DIF were closely scrutinized to determine any inferences concerning the underlying construct and the implications of the sample in which it has been detected [41]. Teresi [44] suggested the use of qualitative analysis (e.g. panel of experts) before an instrument is developed, or if possible, after DIF has been identified in existing measures; thus expert review of the items exhibiting DIF was implemented.

The experts provided detailed knowledge about the construct. Experts determined what could be understood from items exhibiting DIF. A more complete discussion of DIF may be found in elsewhere in this edition [55] and other sources [36, 37, 43, 44, 50, 51]. All IRM analyses were performed using ConQuest [49].

Results

One hundred and one students (out of 1578) did not complete any of the FVSEQ items and were excluded from analyses. Gender was almost evenly distributed with 47.2% boys and 52.8% girls. Ethnicity was assessed via school roster and included 17.4% White, 45.6% Black, 30.1% Hispanic and 6.9% Other students. No demographic differences were observed between students with complete and incomplete data.

CTT item analysis (see Table I) yielded difficulty estimates (item means) between 3.3 (SD = 1.6) and 4.5 (SD = 1.0) based on the scale of 1–5. These values were clearly above the midpoint (midpoint = 3, 'not sure'), indicating that on average the responses were not as difficult to agree with. The CITC were acceptable to high (0.34 to 0.60). Internal consistency was excellent (Cronbach's $\alpha = 0.90$) [26, 39].

The scree plot criterion with principle axis factoring confirmed the FVSEQ included one dominant factor. The percentage of variance explained was 29.7 and 9.8% for the first and second factor, respectively. Comparison of model fit of the RSM and PCM involved using the LRT and comparing the item fit indices [31]. LRT for the deviances between the two models was significant (LRT = 363.10, df = 69, $P < 0.0001$), indicating that the PCM estimates significantly improved the model fit. However, LRT is influenced by sample size; therefore, the nature of the misfit was also examined. Examination of the RSM item difficulty estimates yielded significant misfit of the category step estimates. Although all (100%) average item infit indices were within the acceptable range, all (100%) category step indices exhibited misfit. All (100%) infit indices for the average item estimates for the PCM (not shown) were within the range of acceptable fit and only 3.1% of the item step indices exhibited misfit. Therefore, the PCM was considered the superior model and was accepted as the final model for the data.

PCM item and person estimates are shown in the Wright map (Fig. 1). The participant SE (ability) estimates and the item difficulty (threshold) estimates are on the same logit scale. The logit scale is indicated in the outermost left column. The next section contains the ability estimates that are linked to the column on the right containing the item difficulty estimates. In an ideal situation, the ability distribution would be normally distributed from –3 to +3 and item difficulty estimates would span the entire range of ability. As shown in the figure, the difficulty of the items did not target persons with high levels of FVSE ability. There were participants with high levels of SE (logits > 1.0); however, there are no items that had difficulty estimates in that area. The restricted range in coverage between both the SE ability distribution and the item thresholds indicated the FVSEQ scale's representation of the SE ability construct was skewed. The lowest item by step threshold (1) for 12 items (1–7, 9, 12, 22–24) and several second step thresholds (1–3, 9, 22) were not targeting persons in the lowest levels in SE. The majority of the first set of items (1–7) dealt with tasks such as 'asking', 'shopping' and 'writing' on a list. The last three items (22, 23, 24) were overall tasks for meeting the recommended number of daily FJV servings. The first step for the remaining items which actually covered the lowest SE ability range consisted of items specific to breakfast, lunch, dinner and snacks, and were mostly 'replacement' items such as eating fruit instead of cookies or candy. The conditional reliability, displayed in Fig. 2, showed acceptable reliability for the FVSEQ except at the extreme ends of SE ability scale.

The PCM IRFs for Item 1 are shown in Fig. 3. This pattern was similar in the remaining items in that the response functions for 'disagree a little' and 'agree a little' never had the highest probability of being selected. For some items, the response option 'not sure' had the highest probability of being

Table I. *Item description, item difficulty, item difficulty by ethnicity and difficulty differences between ethnic groups*

Item description		CTT		IRM					
		Mean	CITC	Item difficulty by ethnicity			Difficulty differences between ethnicity[a]		
				White	Black	Hispanic	W − B[b]	W − H[c]	B − H[d]
9.	At breakfast ... drink a glass of favorite J.	4.5 (1.0)	0.39	−0.19	0.07	0.12			
2.	... ask someone in family to buy favorite F or V.	4.4 (1.0)	0.41	−0.32	0.16	0.15	*−0.48*	*−0.47*	
3.	... go shopping with family for favorite F or V.	4.4 (1.1)	0.45	−0.05	−0.06	0.11			
22.	... eat 2 or more servings of F or J each day.	4.3 (1.1)	0.43	−0.20	0.13	0.06	−0.33	−0.26	
7.	... ask someone in family to have Fs & Js out where I can reach them.	4.3 (1.1)	0.40	−0.19	0.08	0.11		−0.30	
1.	... write favorite F of V on the family's shopping list.	4.2 (1.1)	0.36	0.10	−0.07	−0.03			
12.	For lunch at school, ... eat a F that's served.	4.2 (1.2)	0.48	0.23	−0.07	−0.16		0.38	
6.	... ask someone in family to serve favorite F at dinner.	4.1 (1.2)	0.45	−0.24	0.11	0.13		−0.37	
5.	... ask someone in family to make favorite V dish for dinner.	4.1 (1.2)	0.46	−0.07	0.03	0.05			
4.	... pick out favorite F or V at the store & put it in the shopping basket.	4.0 (1.3)	0.34	−0.15	0.03	0.11		−0.26	
8.	... ask someone in family to have V sticks where I can reach them.	4.1 (1.3)	0.52	−0.11	0.11	0.01			
10.	At breakfast ... add favorite F to favorite cereal.	4.0 (1.4)	0.41	−0.02	0.01	0.01			
23.	... eat 3 or more servings of Vs each day.	3.9 (1.3)	0.52	0.06	−0.02	−0.04			
15.	For snack, ... choose favorite F instead of favorite cookie.	4.0 (1.4)	0.56	0.11	−0.06	−0.05			
14.	For lunch at home, ... eat favorite F instead of usual dessert.	3.9 (1.4)	0.6	0.14	−0.08	−0.06	0.22		
11.	For lunch at school, ... eat a V that's served.	3.9 (1.4)	0.52	0.17	−0.07	−0.10		0.27	
24.	... eat 5 or more servings of Fs & Vs each day.	3.7 (1.4)	0.46	0.08	−0.05	−0.03			
16.	For snack, ... choose favorite F instead of favorite candy bar.	3.8 (1.5)	0.60	0.10	−0.01	−0.09			
13.	For lunch at home, ... eat carrot/celery sticks instead of chips.	3.8 (1.5)	0.52	0.04	0.02	−0.06			
21.	For dinner or supper, ... eat favorite F instead of usual dessert.	3.8 (1.5)	0.55	0.09	−0.07	−0.02			
19.	For snack, ...choose favorite raw V & dip instead of chips.	3.4 (1.5)	0.57	0.12	−0.03	−0.09		0.21	
20.	For dinner or supper, ... eat a casserole with Vs.	3.4 (1.6)	0.46	0.09	−0.07	−0.02			
18.	For snack, ... choose favorite raw V & dip instead of favorite candy bar.	3.3 (1.6)	0.57	0.14	−0.06	−0.08		0.21	
17.	For snack, ... choose favorite raw V & dip instead of favorite cookie.	3.3 (1.6)	0.58	0.07	−0.03	−0.04			

White (W), Black (B), Hispanic (H), 'I am sure I can' (...), fruit (F), juice (J), vegetable (V). Note that differences are displayed only for items exhibiting significant DIF and all infit statistics (not shown) within acceptable range (0.75–1.33).
[a]Small effect (difference < 0.426), moderate effect (italic; 0.426 < difference < 0.638), large effect (difference > 0.638).
[b]Negative value, easier for Whites; positive value, easier for Blacks. [c]Negative value, easier for Whites; positive value, easier for Hispanics. [d]No small, moderate or large effects were observed.

Logit	Latent Ability Distribution		Item Threshold Distribution
2	Extremely High FVSE	|	Extremely difficult to endorse response
		X |	
		XX |	
		X |	
		X |	
		XX |	
		XX |	
		XX |	
		XX |	
		XXX |	
		XXXX |	
		XXXX |	
		XXXX |	
1		XXXXX |	17.4 18.4
		XXXXXX |	19.4 20.4
		XXXXXX |	24.4
		XXXXXXX |	
		XXXXXX |	
		XXXXXX |	11.4 21.4 23.4
		XXXXXX |	4.4 5.4 13.4 14.4 15.4 16.4 17.3 18.3 20.3
		XXXXXXXX |	1.4 6.4 8.4 19.3
		XXXXXXXX |	10.4
		XXXXXXXXXX |	17.2 18.2 21.3 24.3
		XXXXXXXX |	12.4 13.3 16.3 19.2 20.2
		XXXXXXX |	3.4 7.4 11.3 14.3 15.3 22.4 23.3
		XXX |	2.4 4.3 8.3 10.3 13.2 16.2 17.1 18.1 21.2
0		XXX |	1.3 5.3 6.3 19.1 20.1
		XXX |	9.4 11.2 12.3 14.2 15.2
		X |	3.3 7.3 8.2 10.2 13.1 16.1 22.3 24.2
		X |	4.2 11.1 12.2 21.1 23.2
		X |	2.3 5.2 6.2 9.3 10.1 14.1 15.1
		X |	7.2 8.1
		|	1.2 3.2 12.1 22.2 23.1 24.1
		|	4.1 9.2
		|	2.2 3.1 5.1 6.1 7.1
		|	1.1
		|	2.1 9.1 22.1
		|	
-1		|	
		|	
		|	
		|	
		|	
		|	
		|	
		|	
		|	
		|	
		|	
-2	Extremely Low FVSE	|	Extremely easy to endorse response

Fig. 1. Wright map of item thresholds for FVSE (each 'X' represents 11.4 cases and the labels for thresholds show the levels of item and step, respectively).

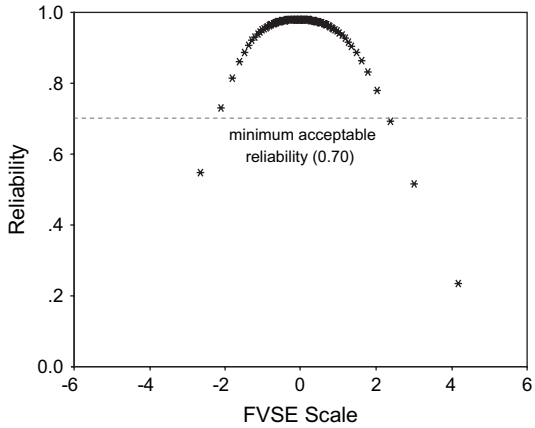

Fig. 2. The reliability for the PCM plotted against the ability estimate for FVSE.

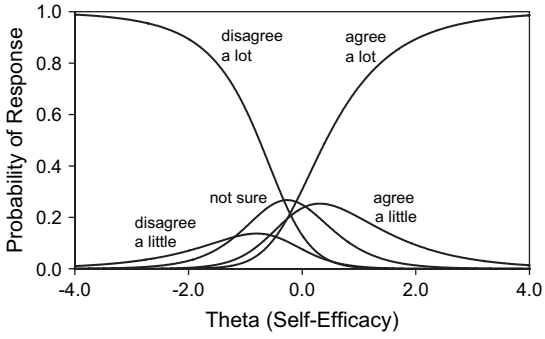

Fig. 3. Response characteristic curves for Item 1, showing the probability of selecting each response at a given level of SE. Note that the response patterns were similar for the remaining items.

selected only among a severely restricted range of SE ability. For the majority of items, the 'not sure' option never had the highest probability of being selected in most items. As shown by the SE curves, the five-point item response format was not fully utilized.

DIF analyses across gender groups did not yield significant DIF (gender by item interaction χ^2 = 18.121, df = 23, P = 0.751) for overall differences for gender. The overall difference between males and females was 0.01 logits. However, DIF across ethnic groups yielded a significant (χ^2 = 122.581, df = 46, $P < 0.001$) group by item interaction (DIF) as well as a significant (χ^2 = 11.690, df = 2, P = 0.003) overall group main effect (impact). The average item difficulty estimates, item difficulty estimates by ethnic group and differences in item difficulty between ethnic groups are displayed in Table I. Overall, the mean SE ability for Whites (0.126) was significantly higher than the mean SE ability for Blacks (0.062) and Hispanics (0.065). Significant DIF existed among 10 items. The magnitude of DIF, however, was small for nine items (Items 4, 6, 7, 11, 12, 14, 18, 19, 22) and moderate for one item (Item 2).

Discussion

The aim of this study was to evaluate the psychometric properties of the FVSEQ items and their stability across gender and ethnic groups using IRM [29]. CTT results showed that the FVSEQ scale had high internal consistency, the item difficulties were moderately easy to difficult and the items were discriminating. IRM complemented the results from CTT by showing (i) the scale yielded adequate reliability except for participants with very low or high levels of SE, (ii) the scale did not appear to adequately assess participants with high levels of SE and (iii) the five-point response format was not fully utilized. Additional information provided by IRM showed no gender DIF, but several items exhibited DIF across ethnic groups. Whites found it significantly easier than Blacks and Hispanics to perform 'asking' tasks, while Blacks and Hispanics found it significantly easier than Whites to perform intake behaviors such as eating a vegetable that is served. Examination by experts suggested this difference was more likely due to real ethnic differences in factors such as family structure, cultural differences in the activity or family dynamics. For example, Cullen *et al.* [28] showed that Hispanics were in families with a permissive style of parenting and Whites were in families with greater meal planning practices. This DIF likely provided additional insight into tailoring interventions to ethnic groups to incorporate cultural differences in family

dynamics. More work is necessary to clarify ethnic differences.

Results from IRM provided more in-depth information about the scale and identified areas in which the FVSEQ could be improved. For example, to target participants with high levels of SE, an item of average difficulty such as 'I can eat a vegetable that is served for school lunch' might be modified for greater difficulty as 'I can eat any vegetable that is served for school lunch every day'. Newer possibilities need to be explored, e.g. 'I can make myself learn to like any vegetable' or 'I can problem solve to overcome any barrier to eating more vegetables'. Also, the current five-point scale should be replaced with fewer response options, or the response options should be changed so that responses will be more uniform.

The strengths of this study include (i) ample sample size necessary to perform IRM, and more specifically, DIF analyses and (ii) the use of an existing and previously validated instrument to measure FVSE. The limitations of this study include the disregard of possible clustering within schools and the determination of ethnicity via school roster instead of self-report. Additional limitations include the use of 'not sure' as a neutral category, i.e. some error in estimating the true level of SE may be associated with the assumption that those who are 'not sure' reflect more positivity than 'disagree'.

In summary, IRM provided (i) difficulty estimates that were not dependent on this sample, (ii) the ability estimates of FV SE were not specific to the items on the instrument and (iii) measurement error was a function of ability. The major practical value of applying IRM was that although the test was adequately reliable, it was not measuring the full range of the construct. The scale was not able to provide discrimination among participants with higher levels of SE, thus indicating a need to revise the instrument. Although some items exhibited minor DIF, this can probably be ignored. Closer examination of these items in future studies should be performed to ensure that the items are not biased. A questionnaire revised to assess the full range of SE difficulty estimates should correlate better with FV intake. This work remains to be done.

Acknowledgements

This research was largely funded by a grant from the National Institutes of Health, grant R01 CA-75614. This work is also a publication of the US Department of Agriculture (USDA/ARS) Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, and Houston, Texas, funded in part with federal funds from the USDA/ARS under Cooperative Agreement No. 58-6250-6001. The contents of this publication do not necessarily reflect the views or policies of the USDA, nor does mention of trade names, commercial products or organizations imply endorsement from the US government.

Conflict of interest statement

None declared.

References

1. Center for Disease Control and Prevention. Guidelines for school health programs to promote lifelong healthy eating. *Morb Mortal Wkly* 1996; **45**: 1–33.
2. US Department of Health and Human Services, Centers for Disease Control and Prevention. *Assessing Health Risk Behaviors Among Young People: Youth Risk Behavior Surveillance System 2004*. Atlanta, GA: National Center for Chronic Disease Prevention and Health Promotion, 2004.
3. US Department of Health and Human Services, Centers for Disease Control and Prevention. *Physical Activity and Good Nutrition: Essential Elements to Prevent Chronic Disease and Obesity 2003*. National Center for Chronic Disease Prevention and Health Promotion, 2003.
4. Bandura A. *Foundations for Thoughts and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice Hall, 1986.
5. Reynolds KD, Yaroch AL, Franklin FA *et al*. Testing mediating variables in a school-based nutrition intervention program. *Health Psychol* 2002; **21**: 51–60.
6. Strecher VJ, DeVellis BM, Becker MH *et al*. The role of self-efficacy in achieving health behavior change. *Health Educ Q* 1986; **13**: 73–92.
7. Baranowski T, Davis M, Resnicow K *et al*. Gimme 5 fruit, juice, and vegetables for fun and health: outcome evaluation. *Health Educ Behav* 2000; **27**: 96–111 [Erratum appears in *Health Educ Behav* 2000; **27**: 390].
8. Edmundson E, Parcel GS, Feldman HA *et al*. The effects of the Child and Adolescent Trial for Cardiovascular Health upon psychosocial determinants of diet and physical activity behavior. *Prev Med* 1996; **25**: 442–54.

9. Long JD, Stevens KR. Using technology to promote self-efficacy for healthy eating in adolescents. *J Nurs Scholarsh* 2004; **36**: 134–9.
10. Neumark-Sztainer D, Wall M, Perry C *et al.* Correlates of fruit and vegetable intake among adolescents. Findings from Project EAT. *Prev Med* 2003; **37**: 198–208.
11. Resnicow K, Davis-Hearn M, Smith M *et al.* Social-cognitive predictors of fruit and vegetable intake in children. *Health Psychol* 1997; **16**: 272–6.
12. Saksvig BI, Gittelsohn J, Harris SB et *al.* A pilot school-based healthy eating and physical activity intervention improves diet, food knowledge, and self-efficacy for native Canadian children. *J Nutr* 2005; **135**: 2392–8.
13. Stevens J, Story M, Ring K *et al.* The impact of the Pathways intervention on psychosocial variables related to diet and physical activity in American Indian schoolchildren. *Prev Med* 2003; **37**: S70–9.
14. Wilson DK, Friend R, Teasley N *et al.* Motivational versus social cognitive interventions for promoting fruit and vegetable intake and physical activity in African American adolescents. *Ann Behav Med* 2002; **24**: 310–9.
15. Domel SB, Baranowski T, Davis HC *et al.* Psychosocial predictors of fruit and vegetable consumption among elementary school children. *Health Educ Res Theory Pract* 1996; **11**: 299–308.
16. Parcel GS, Simons-Morton B, O'Hara NM *et al.* School promotion of healthful diet and physical activity: impact on learning outcomes and self-reported behavior. *Health Educ Q* 1989; **16**: 181–99.
17. Granner ML, Sargent RG, Calderon KS *et al.* Factors of fruit and vegetable intake by race, gender, and age among young adolescents. *J Nutr Educ Behav* 2004; **36**: 173–80.
18. Hagler AS, Norman GJ, Radick LR *et al.* Comparability and reliability of paper- and computer-based measures of psychosocial constructs for adolescent fruit and vegetable and dietary fat intake. *J Am Diet Assoc* 2005; **105**: 1758–64.
19. Zabinski MF, Daly T, Norman GJ *et al.* Psychosocial correlates of fruit, vegetable, and dietary fat intake among adolescent boys and girls. *J Am Diet Assoc* 2006; **106**: 814–21.
20. Sallis JF, Pinski RB, Grossman RM *et al.* The development of self-efficacy scales for health related diet and exercise behaviors. *Health Educ Res* 1988; **3**: 283–92.
21. Cusatis DC, Shannon BM. Influences on adolescent eating behavior. *J Adolesc Health* 1996; **18**: 27–34.
22. Monge-Rojas R, Nunez HP, Garita C *et al.* Psychosocial aspects of Costa Rican adolescents' eating and physical activity patterns. *J Adolesc Health* 2002; **31**: 212–9.
23. Mangunkusumo RT, Duisterhout JS, de Graaff N *et al.* Internet versus paper mode of health and health behavior questionnaires in elementary schools: asthma and fruit as examples. *J Sch Health* 2006; **76**: 80–6.
24. Reynolds KD, Killen JD, Bryson SW *et al.* Psychosocial predictors of physical activity in adolescents. *Prev Med* 1990; **19**: 541–51.
25. Vereecken CA, Van Damme W, Maes L. Measuring attitudes, self-efficacy, and social and environmental influences on fruit and vegetable consumption of 11- and 12-year-old children: reliability and validity. [See comment.] *J Am Diet Assoc* 2005; **105**: 257–61.
26. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart, and Winston, Inc., 1986.
27. Backman DR, Haddad EH, Lee JW *et al.* Psychosocial predictors of healthful dietary behavior in adolescents. *J Nutr Educ Behav* 2002; **34**: 184–92.
28. Cullen KW, Baranowski T, Owens E *et al.* Ethnic differences in social correlates of diet. *Health Educ Res* 2002; **17**: 7–18.
29. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response Theory*. Thousand Oaks, CA: Sage Publications, Inc., 1991.
30. Wilson M, Allen D, Li JC. Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. *Health Educ Res* 2006; **21**(Suppl 1): i4–i18.
31. Wilson M, Allen D, Li JC. Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. *Health Educ Res* 2006; **21**(Suppl 1): i19–i32.
32. Lord F. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.
33. Lord F, Novick MR. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968.
34. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press, 1980.
35. Masse LC, Dassa C, Gauvin L *et al.* Emerging measurement and statistical methods in physical activity research. *Am J Prev Med* 2002; **23**: 44–55.
36. Swaminathan H, Rogers H. Detecting differential item functioning using logistic regression procedures. *J Educ Meas* 1990; **27**: 361–70.
37. Wilson M. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Associates, 2005.
38. Baranowski T, Baranowski J, Cullen KW *et al.* Squire's Quest! Dietary outcome evaluation of a multimedia game. *Am J Prev Med* 2003; **24**: 52–61.
39. Nunnally JC, Bernstein IH. *Psychometric Theory*. New York: McGraw-Hill, 1994.
40. SPSS Inc. *SPSS for Windows Release 11.0.1*. Chicago, IL: SPSS Inc., 2001.
41. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 2001.
42. Reise SP, Yu J. Parameter recovery in the graded response model using MULTILOG. *J Educ Meas* 1990; **27**: 133–44.
43. Orlando M, Marshall GN. Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. *Psychol Assess* 2002; **14**: 50–9.
44. Teresi J. Differential item functioning and health assessment. In: *Conference on Improving Health Outcomes Assessment Based on Modern Measurement Theory and Computerized Adaptive Testing*. Bethesda, MD, 2004.
45. Bolt DM. A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Appl Meas Educ* 2002; **15**: 113–41 .

46. Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; **43**: 561–73.
47. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982; **47**: 149–74.
48. Wu M. Plausible values. *Rasch Meas Trans* 2004; **18**: 976–78.
49. Wu M, Adams R, Haldane S. *ConQuest*. Berkeley, CA: Australian Council for Educational Research University of California, 2003.
50. Shepard LA. Definition of bias. In: Berk RA, Baltimore MD (eds). *Handbook of Methods for Detecting Test Bias*. Baltimore, MD: Johns Hopkins University Press, 1982, 9–30.
51. Smith RM. Detecting item bias with the Rasch model. In: Smith EV, Smith RM (eds). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press, 2004, 391–418.
52. Orlando M. Critical issues to address when applying item response theory (IRT) models. In: *Conference on Improving Health Outcomes Assessment Based on Modern Measurement Theory and Computerized Adaptive Testing*. Bethesda, MD: Hyatt, 2004.
53. Paek I. Investigation of differential item function: comparisons among approaches, and extension to a multidimensional context. *Unpublished PhD Dissertation*. Berkeley, CA: University of California, 2002.
54. Angoff WH. Perspectives on differential item functioning methodology. In: Holland PW, Wainer H (eds). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum and Associates, 1993, 3–24.
55. Baranowski T, Allen D, Masse LC *et al*. Does participation in an intervention affect the responses on self-report questionnaires? *Health Educ Res* 2006; **21**(Suppl 1): i98–i109.

Received on February 16, 2006; accepted on September 25, 2006